

# Mutual-Information Optimized Quantization for LDPC Decoding of Accurately Modeled Flash Data

Yiadong Wang, Guiqiang Dong, Tong Zhang and Richard Wesel  
 wjd@ee.ucla.edu, dongguiqiang@gmail.com, tzhang@ecse.rpi.edu, wesel@ee.ucla.edu

**Abstract**—High-capacity NAND flash memories use multi-level cells (MLCs) to store multiple bits per cell and achieve high storage densities. Higher densities cause increased raw bit error rates (BERs), which demand powerful error correcting codes. Low-density parity-check (LDPC) codes are a well-known class of capacity-approaching codes in AWGN channels. However, LDPC codes traditionally use soft information while the flash read channel provides only hard information. Low resolution soft information may be obtained by performing multiple reads per cell with distinct word-line voltages.

We select the values of these word-line voltages to maximize the mutual information between the input and output of the equivalent multiple-read channel under any specified noise model. Our results show that maximum mutual-information (MMI) quantization provides better soft information for LDPC decoding given the quantization level than the constant-pdf-ratio quantization approach. We also show that adjusting the LDPC code degree distribution for the quantized setting provides a significant performance improvement.

## I. INTRODUCTION

Flash memory is a low-power non-volatile device that can carry a large amount of data within a small area. The original NAND flash architecture was called single-level-cell (SLC) flash and used only one nonzero charge level to store one bit. More recent devices use multiple levels and are referred to as multiple-level cell (MLC) flash. Four and eight levels per cell can be found in use, and the number of levels will increase further to provide more storage capability [1][2].

The increase in the number of levels and aggressive feature-size reduction cause cell-to-cell interference and retention noise to become more severe than for the original SLC flash memories [3]. Powerful codes are required to cope with these obstacles and maximize the potential of the system.

Low-density parity-check (LDPC) codes are a class of capacity-approaching codes for the AWGN channel [4]. Flash systems typically only provide hard reliability information after the reading process while LDPC codes typically utilize soft reliability information. This paper uses realistic channel models to demonstrate that efficiently extracting soft information with a few extra reads in the cell can significantly improve LDPC code performance in flash memory.

Our previous analysis [5] used pulse-amplitude modulation (PAM) with Gaussian noise to model Flash cell threshold voltage levels. We investigated how to select the word-line voltages to maximize the mutual information between the input and the output of the equivalent read channel. With

This research was supported by a gift from Inphi Corp. and UC Discovery Grant 192837.

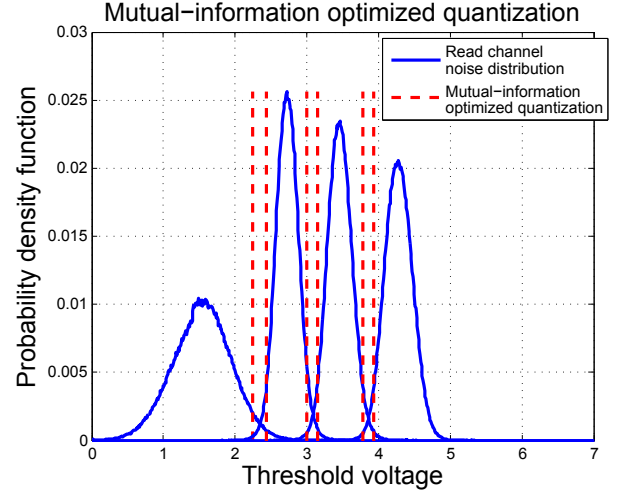


Fig. 1: Mutual-information optimized quantization for the six-month data.

carefully selected word-line voltage for each of the reads, we represent the multiple-read channel as a probability transition matrix and decode the data using a standard belief-propagation algorithm. The maximum mutual information (MMI) approach is also explored in [6] [7] for the design of the message-passing decoders of LDPC codes to optimize the quantization of the binary-input channel output.

This paper extends the analysis in [5] to any noise model for the flash memory read channel. As an example, we model a four-level six-read MLC as a four-input seven-output discrete channel. Instead of assuming Gaussian noise distributions as in [5], we numerically compute the probability transition matrix using the retention noise model from [8].

Fig. 1 shows the four conditional threshold-voltage probability density functions generated according to [8] and the resulting six MMI word-line voltages after six months retention time. While the conditional noise for each transmitted (or written) threshold voltage is similar to that of a Gaussian, the variance of the conditional distributions varies greatly across the four possible threshold voltages. Note that the lowest threshold voltage has by far the largest variance.

In [9], a heuristic quantization algorithm sets the word-line voltages to the value where the two adjacent probability density functions have a constant ratio  $R$ . This paper compares the MMI approach with the constant-ratio method of [9] using the realistic channel model of [8] and shows that the MMI approach generally outperforms the constant-ratio method.

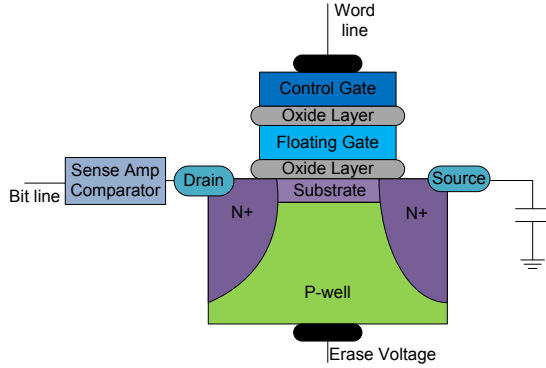


Fig. 2: A NAND flash memory cell.

This paper also explores how the quantized setting should be considered in the selection of the LDPC degree distribution. LDPC codes are usually designed with the degree distribution optimized for the AWGN channel [4]. However, our simulations show that, in the quantized setting, adjusting this “optimal” degree distribution can significantly improve performance.

Section II introduces the basics of the NAND flash memory model and LDPC codes. Section III describes the retention noise model using MLC as an example and shows how to obtain word-line voltages by maximizing the mutual information of the equivalent read channel. Section IV provides simulation results demonstrating the benefits adjusting the degree distribution to the quantized setting and compares the MMI approach with the constant ratio method for extracting soft information. Section V delivers the conclusions.

## II. BACKGROUND

This section introduces the basics of NAND flash memory and LDPC codes.

### A. Basics of NAND Flash Memory

This paper focuses on the NAND-architecture flash memory that is currently the most prevalent architecture. In the NAND architecture, each memory cell features a transistor with a control gate and a floating gate. To store information, a relatively large voltage is applied to the control gate, which adds a specified amount of charge to the floating gate through Fowler-Nordheim tunneling [10].

Fig. 2 shows the configuration of a NAND flash memory cell. To read a memory cell, the charge level written to the floating gate is detected by applying a specified word-line voltage to the control gate and measuring the transistor drain current. The drain current is compared to a threshold by a sense amp comparator. If the drain current is above the comparator threshold, then the word-line voltage was sufficient to turn on the transistor, indicating that the charge written to the floating gate was insufficient to prevent the transistor from turning on. If the drain current is below the threshold, the charge added to the floating gate was sufficient to prevent the applied word-line voltage from turning on the transistor.

The sense amp comparator only provides one bit of information about the charge level present in the floating gate. The word-line voltage required to turn on a particular transistor (called the threshold voltage) can vary from cell to cell for a variety of reasons. For example, the floating gate can receive extra charge when nearby cells are written, the floating gate can be overcharged during the write operation, or the floating gate can lose charge due to leakage in the retention period [11]. We are not optimizing word-line voltages for a particular cell, but rather for the threshold voltage distribution over all cells at a certain retention time.

In [5], we assumed an i.i.d. Gaussian threshold voltage for each level of an MLC flash memory cell. More precise models such as the model in [11] in which the lowest and highest threshold voltage distributions have a higher variance and the model in [12] in which the lowest threshold voltage (the one associated with zero charge level) is Gaussian and the other threshold voltages have Gaussian tails but a uniform central region are sometimes used. The model in [8] is similar to [12], but is derived by explicitly accounting for real dominating noise sources, such as inter-cell interference, program injection statistics, random telegraph noise and retention noise. In this paper, we use the read channel model of [8].

### B. Basics of LDPC codes

LDPC codes are well-known as capacity-approaching codes of the AWGN channel, and they are defined by sparse parity-check matrices. The degree distribution of LDPC codes can be optimized to operate closely to the capacity of an AWGN channel [4]. For a given degree distribution, we can generate LDPC codes using several algorithms, such as the PEG algorithm [14], and the ACE algorithm [13].

Storage systems typically require frame-error-rates lower than  $10^{-15}$ , making the design of LDPC codes with low error-floors necessary for applications to flash memory. This topic has generated a significant amount of recent research including [15] [16] [17] [18] [19].

Iterative belief propagation algorithms are used for decoding LDPC codes. Soft reliability information at the receiver usually can significantly improve the performance of belief-propagation decoders. Conversely, a coarse quantization of the received information can degrade the performance of an LDPC code.

The remainder of this paper presents a general quantization approach to select word-line voltages to maximize the mutual information for an  $N$ -level  $M$ -read Flash memory and any noise model. These word-line voltages are then used to gather quantized soft information for an LDPC decoder. We also compare the MMI approach with an alternative quantization approach and show how adjusting the degree distribution can improve performance in a quantized setting.

## III. QUANTIZATION FOR MLC FLASH

For any  $N$ -level  $M$ -read Flash memory and any noise model, the multiple-read channel can be represented by a probability transition matrix after choosing the word-line voltage

for each of the reads, and the data can be decoded with a standard belief-propagation algorithm. The equivalent discrete channel induces a mutual information between the  $N$  inputs and  $M + 1$  outputs.

#### A. Maximum Mutual Information (MMI) Quantization

The approach is to select the set of word-line voltages, which maximizes that mutual information.

This section uses a 4-level 6-read MLC with retention noise as an example. For 4-level MLC flash memory, each cell can store 2 bits of information. Gray labeling (00, 01, 11, 10) minimizes the raw bit error rate for these four levels. In 4-level MLC flash, each cell is typically compared to three word-line voltages and thus the output of the comparator has four distinct quantization regions. In this section, we consider three additional word-line voltages (for a total of six) and quantize the threshold voltage to seven distinct regions as shown in Fig. 1. Fig. 3 presents the post-quantization channel model, a 4-input 7-output discrete memoryless channel.

Suppose the 6 word-line voltages are  $q_1, q_2, \dots, q_6$ . We can numerically compute the probability transition matrix using the probability density function generated from the retention noise model in [8]. Fig. 1 shows the probability density function generated from [8] and the resulting MMI word-line voltages after 6 months retention time.

Since the retention noise model itself is not an analytic expression and certainly not symmetric, we need to numerically compute all the probabilities in Fig. 3 and calculate the mutual information between the input and output:

$$\begin{aligned}
& I(X; Y) \\
&= H(Y) - H(Y|X) \\
&= H\left(\frac{p_{11} + p_{21} + p_{31} + p_{41}}{4}, \frac{p_{12} + p_{22} + p_{32} + p_{42}}{4}, \right. \\
&\quad \frac{p_{13} + p_{23} + p_{33} + p_{43}}{4}, \frac{p_{14} + p_{24} + p_{34} + p_{44}}{4}, \\
&\quad \frac{e_{1a} + e_{2a} + e_{3a} + e_{4a}}{4}, \frac{e_{1b} + e_{2b} + e_{3b} + e_{4b}}{4}, \\
&\quad \left. \frac{e_{1c} + e_{2c} + e_{3c} + e_{4c}}{4}\right) \\
&\quad - \frac{1}{4}H(p_{11}, p_{12}, p_{13}, p_{14}, e_{1a}, e_{1b}, e_{1c}) \\
&\quad - \frac{1}{4}H(p_{21}, p_{22}, p_{23}, p_{24}, e_{2a}, e_{2b}, e_{2c}) \\
&\quad - \frac{1}{4}H(p_{31}, p_{32}, p_{33}, p_{34}, e_{3a}, e_{3b}, e_{3c}) \\
&\quad - \frac{1}{4}H(p_{41}, p_{42}, p_{43}, p_{44}, e_{4a}, e_{4b}, e_{4c}). \tag{1}
\end{aligned}$$

The mutual information in (1) is in general not a quasi-concave function in terms of the word-line voltages  $q_1, q_2, \dots, q_6$ , although it is quasi-concave for the simple model of two symmetric Gaussians with symmetric word-line voltages studied in [5]. Since (1) is a continuous and smooth function and locally quasi-concave in the range of our interest, we can numerically compute the maximum mutual information with a careful use of bisection search.

After optimizing the word-line voltages  $q_1, q_2, \dots, q_6$ , the equivalent 4-input 7-output discrete channel has the maximum

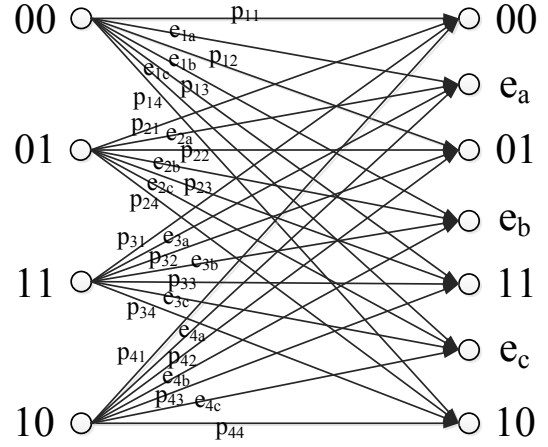


Fig. 3: Quantization model for 4-MLC with 6 reads.

mutual information between the input and output. We can easily extend this technique to any other  $N$ -level  $M$ -read Flash memory and any other noise model with known conditional distributions.

#### B. Constant PDF-Ratio Quantization

In [9], a quantization algorithm sets each word-line voltage to the value where the two adjacent pdfs have a constant ratio  $R$ . The difficult step in this algorithm is the selection of  $R$ , which is accomplished by a heuristic process of simulation and adjustment. The goal of simulation and adjustment is to optimize the decoder performance.

Fig. 4 compares the mutual information obtained by using the MMI approach and the constant ratio method for a variety of  $R$  values. We note that with certain  $R$  values, the constant ratio method can provide mutual information that is close to the maximum. However, finding the best  $R$  can be a challenge.

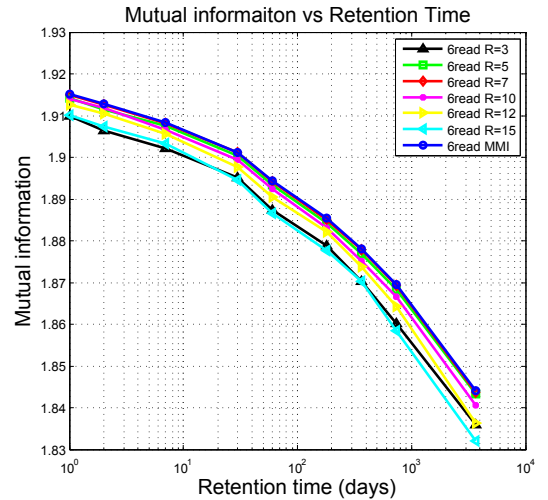


Fig. 4: Mutual information comparison between the MMI approach and constant ratio method with retention data.

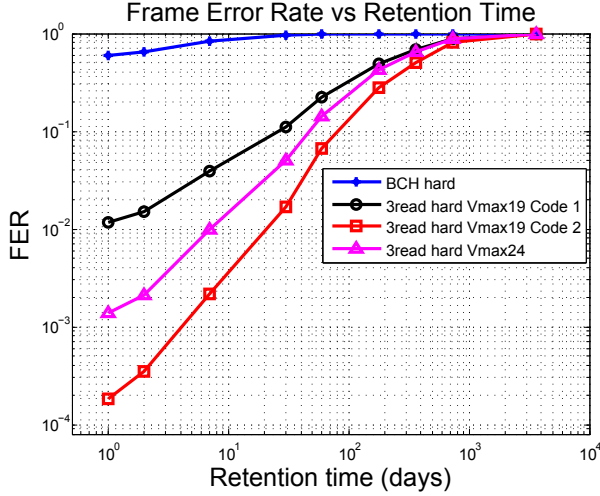


Fig. 5: Simulation results for 4-level MLC using hard quantization.

#### IV. LDPC PERFORMANCE COMPARISONS

This paper uses rate-0.9021 irregular LDPC codes with block length  $n = 9118$  and dimension  $k = 8225$  for simulation. LDPC matrices are constructed according to the degree distributions using the ACE algorithm [13] together with the stopping-set check algorithm [20] to optimize the LDPC matrix. Simulations were performed using a sequential belief propagation decoder (layered belief propagation) [21]. A rate-0.9021 BCH code with block length  $n = 9152$  and dimension  $k = 8256$  provides a baseline for comparison.

##### A. Degree distribution in a quantized setting

Two of the LDPC codes studied feature distinct degree distributions with maximum variable degree 19. For Code 1, the degree distribution is the usual optimal degree distribution for AWGN [4]. For Code 2, the initial AWGN-optimal degree distribution is adjusted to improve performance in a quantized setting as follows:

Hard decoding induces small absorbing sets such as  $(4, 2)$ ,  $(5, 1)$ ,  $(5, 2)$  absorbing sets in Code 1. To preclude these absorbing sets, we increase all the degree-3 variable nodes to have degree 4 to produce Code 2. Code 2 significantly outperforms Code 1 under hard decoding.

We also simulated another code with the maximum variable degree 24 with degree distribution optimized for AWGN (Code 3). Fig. 5 shows frame error rate versus retention time under hard decoding for these three codes. Code 3 has an even better threshold in AWGN than Code 1, but the newly designed Code 2 with the lower AWGN threshold still outperforms it under hard quantization. This demonstrates that a superior AWGN threshold does not necessarily imply superior performance under hard decoding. Of course when simulated in AWGN with full resolution soft decoding, Code 3 performs better than Code 1, and Code 1 performs better than Code 2.

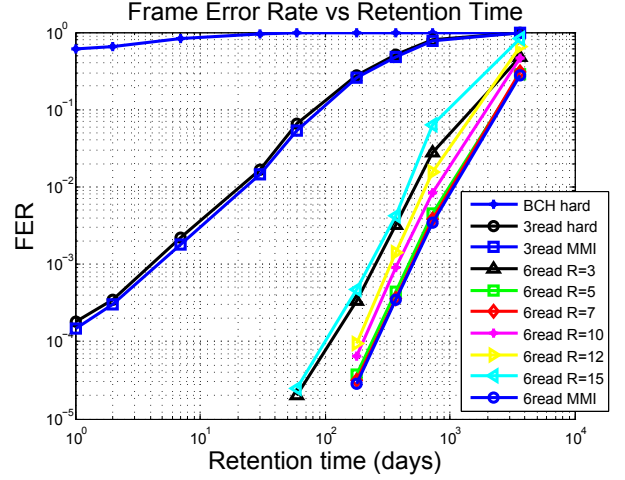


Fig. 6: Simulation results for 4-level MLC using MMI and constant  $R$  with retention data and Code 2.

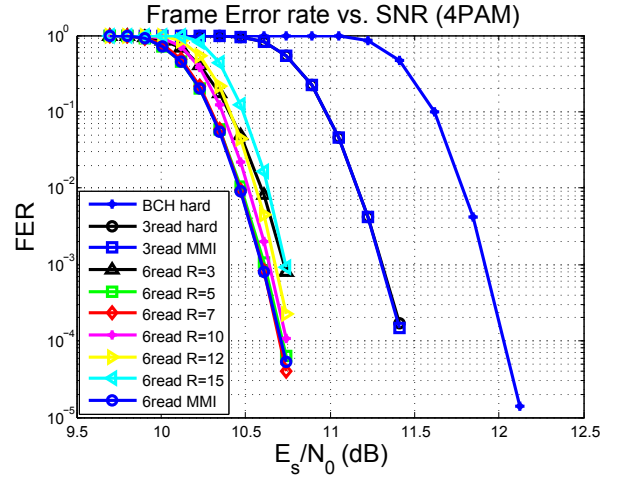


Fig. 7: Simulation results for 4-level MLC using MMI and constant  $R$  with the Gaussian model and Code 2.

##### B. Comparison of quantization methods

Fig. 6 shows frame error rate (FER) plotted versus retention time for Code 2 under a variety of quantizations. This LDPC code has the adjusted degree distribution to lower the FER for hard quantization. Since the noise model is not symmetric, the MMI approach with 3 reads has a slightly larger mutual information than the hard quantization with 3 reads, and thus performs slightly better.

In this plot, the performance of the LDPC codes is closely related to the mutual information of the equivalent channel given in Fig. 4. The  $R = 7$  case has the closest mutual information to the MMI and its performance is also very close to that of the MMI approach. The MMI approach outperforms the constant  $R$  method with most  $R$ s in this plot. For comparison, Fig. 7 shows similar results for six reads using the Gaussian model in [5].

Figs. 8 and 9 are analogous to Figs. 6 and 7 but for Code 1, which was not adjusted for the quantized setting. In this case the constant ratio method with  $R = 15$  slightly



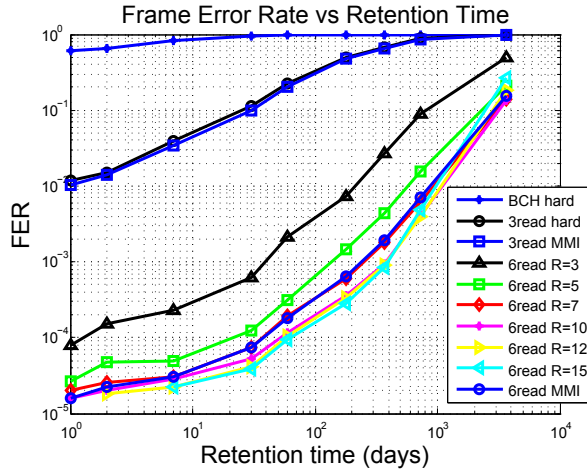


Fig. 8: Simulation results for 4-level MLC using MMI and constant R with retention data and Code 1.

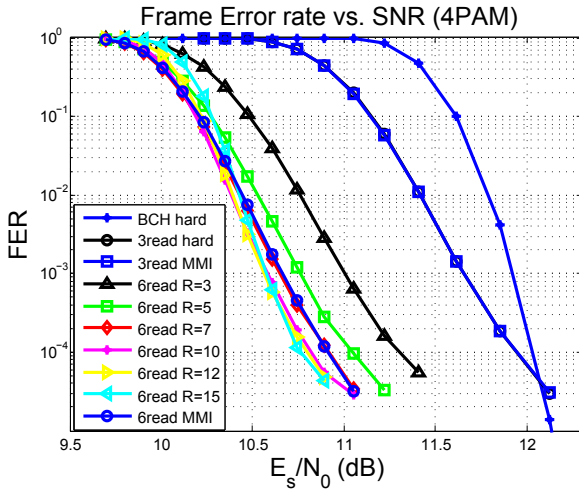


Fig. 9: Simulation results for 4-level MLC using MMI and constant R with the Gaussian model and Code 1.

outperforms the MMI approach. For comparison, Fig. 9 shows similar results for six reads using the Gaussian model. This example reflects our experience that the only cases in which the constant ratio method (slightly) outperforms the MMI method are cases in which the LDPC degree distribution is not well-matched to the channel. In other words, when one has identified a good code for the channel, MMI will give the best quantization. The degree distribution for which MMI was not optimal was also not the best choice of degree distribution.

## V. CONCLUSION

This paper shows that using a small amount of soft information significantly improves the performance of LDPC codes and demonstrates a clear performance advantage over conventional BCH codes. In order to maximize the performance benefit of the soft information, we develop a word-line-voltage-selection method that maximizes the mutual information between the input and output of the DMC equivalent to the quantized read channel. This method can be applied

to any given channel model and provides an effective and efficient estimate of the word-line voltages, as compared to other existing quantization techniques. Possible directions for future research include the design of better high-rate LDPC codes specifically for the flash memory channel, and the analysis of the corresponding error-floor properties.

## REFERENCES

- [1] Y. Li, S. Lee, and et al. A 16 Gb 3b/cell NAND Flash Memory in 56nm With 8MB/s Write Rate. In *Proc. of ISSCC*, pages 506–632, Feb. 2008.
- [2] C. Trinh, N. Shibata, and et al. A 5.6MB/s 64 Gb 4b/Cell NAND Flash Memory in 43nm CMOS. In *Proc. of ISSCC*, page 246, Feb. 2009.
- [3] J.-D. Lee, S.-H. Hur, and J.-D. Choi. Effects of floating-gate interference on NAND flash memory cell operation. *IEEE Electron Device Letters*, 23(5):264–266, May 2002.
- [4] T. Richardson, M. Shokrollahi, and R. Urbanke. Design of capacity-approaching irregular low-density parity-check codes. *IEEE Trans. Inform. Theory*, 47(2):616–637, Feb. 2001.
- [5] J. Wang, T.A. Courtade, H. Shankar, and R.D. Wesel. Soft Information for LDPC Decoding in Flash: Mutual-Information Optimized Quantization. In *Proc. IEEE Global Telecomm. Conf. (GLOBECOM)*, Houston, TX, Dec. 2011.
- [6] J. K.-S. Lee and J. Thorpe. Memory-Efficient Decoding of LDPC Codes. In *Proc. IEEE Int. Symp. on Info. Theory (ISIT)*, Adelaide, Australia, July 2005.
- [7] B. M. Kurkoski and H. Yagi. Quantization of Binary-Input Discrete Memoryless Channels, with Applications to LDPC Decoding. *Submitted to IEEE Trans. Inform. Theory*. Available <http://arxiv.org/abs/1107.5637>.
- [8] Q. Wu, G. Dong, and T. Zhang. Exploiting Heat-Accelerated Flash Memory Wear-Out Recovery to Enable Self-Healing SSDs. In *USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage)*, June 2011.
- [9] G. Dong, N. Xie, and T. Zhang. On the Use of Soft-Decision Error-Correcting Codes in NAND Flash Memory. *IEEE Trans. Circ. and Sys.*, 58(2):429–439, Feb. 2011.
- [10] R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti. Introduction to Flash Memory. *Proc. IEEE*, 91(4), April 2003.
- [11] Y. Maeda and K. Haruhiko. Error Control Coding for Multilevel Cell Flash Memories Using Nonbinary Low-Density Parity-Check Codes. In *24th IEEE Int. Symp. on Defect and Fault Tolerance in VLSI Systems*, Chicago, IL, Oct. 2009.
- [12] S. Li and T. Zhang. Improving Multi-Level NAND Flash Memory Storage Reliability Using Concatenated BCH-TCM Coding. *IEEE Trans. VLSI Systems*, 18(10):1412–1420, Oct. 2010.
- [13] T. Tian, C. Jones, J. D. Villaseñor, and R. D. Wesel. Selective Avoidance of Cycles in Irregular LDPC Code Construction. *IEEE Trans. Comm.*, 52(8):1242–1247, Aug. 2004.
- [14] X.-Y. Hu, E. Eleftheriou, and D.-M. Arnold. Progressive edge-growth Tanner graphs. In *Proc. IEEE GLOBECOM*, San Antonio, TX, Feb. 2001.
- [15] T. Richardson. Error-floors of LDPC codes. In *Proc. 41st Annual Allerton Conf.*, Monticello, IL, Oct. 2003.
- [16] J. Wang, L. Dolecek, and R.D. Wesel. Controlling LDPC Absorbing Sets via the Null Space of the Cycle Consistency Matrix. In *Proc. IEEE Int. Conf. on Comm. (ICC)*, Kyoto, Japan, June. 2011.
- [17] M. Ivkovic, S. K. Chilappagari, and B. Vasic. Eliminating trapping sets in low-density parity-check codes by using Tanner graph covers. *IEEE Trans. Inform. Theory*, 54(8):3763–3768, Aug. 2008.
- [18] D. V. Nguyen, B. Vasic, and M. Marcellin. Structured LDPC Codes from Permutation Matrices Free of Small Trapping Sets. In *Proc. IEEE Info. Theory Workshop (ITW)*, Dublin, Ireland, Sept. 2010.
- [19] Q. Huang, Q. Diao, S. Lin, and K. Abdel-Ghaffar. Cyclic and quasi-cyclic LDPC codes: new developments. In *Proc. Info. Theory and Appl. Workshop*, San Diego, CA, Feb. 2011.
- [20] A. Ramamoorthy and R. D. Wesel. Construction of Short Block Length Irregular LDPC Codes. In *Proc. IEEE Int. Conf. on Comm. (ICC)*, Paris, France, June. 2004.
- [21] E. Yeo, P. Pakzad, B. Nikolic, and V. Anantharam. High Throughput Low-Density Parity-Check Decoder Architectures. In *Proc. IEEE GLOBECOM*, pages 3019–3024, Nov. 2001.